Oei, L., Koromani, F., Breda, S. J., Schousboe, J. T., Clark, E. M., van Meurs, J. B. J., Ikram, M. A., Waarsing, J. H., van Rooij, F. J. A., Zillikens, M. C., Krestin, G. P., Oei, E. H. G., & Rivadeneira, F. (2018). Osteoporotic Vertebral Fracture Prevalence Varies Widely Between Qualitative and Quantitative Radiological Assessment Methods: The Rotterdam Study. *Journal of Bone and Mineral Research*, *33*(4), 560-568. https://doi.org/10.1002/jbmr.3220

Osteoporotic Vertebral Fracture Prevalence Varies Widely Between Qualitative and Quantitative Radiological Assessment Methods: The Rotterdam Study

Oei L[1,2]*, Koromani F[1,2,3]*, Breda SJ[3], Schousboe JT[4], Clark EM[5], van Meurs JBJ[1], Ikram MA[2], Waarsing JH[6], van Rooij FJA[2], Zillikens MC[1], Krestin GP[3], Oei EHG[3]**, Rivadeneira F[1,2]**

[1] *Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands*
[2] *Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands*
[3] *Department of Radiology, Erasmus MC, Rotterdam, The Netherlands*
[4] *Park Nicollet Clinic and HealthPartners Institute, HealthPartners Inc., Minneapolis, MN, USA*
[5] *Musculoskeletal Research Unit, School of Clinical Science, University of Bristol, Southmead Hospital, Bristol, UK*
[6] *Department of Orthopedics, Erasmus MC,*

Keywords: osteoporosis; fracture; vertebral; diagnosis

Word count: Text: 4,117 Abstract: 281 Tables: 5 Figures: 4

Osteoporosis, Epidemiology, Screening, Radiology

Corresponding author:

Fernando Rivadeneira MD, PhD

Departments of Internal Medicine

and Epidemiology

P.O. Box 2040 Ee5-59b

3000CA Rotterdam

The Netherlands

Email: f.rivadeneira@erasmusmc.nl

Phone number: +31 10 7044015

Supplemental data: Tables 2, Figures 2

ABSTRACT

Background: Accurate diagnosis of vertebral osteoporotic fractures is crucial for the identification of individuals at high risk of future fractures. Different methods for radiological assessment of vertebral fractures exist, but a gold standard is lacking.

The aim of our study was to estimate statistical measures of agreement and prevalence of osteoporotic vertebral fractures in the population-based Rotterdam Study, across two assessment methods. Methods: The quantitative morphometry assisted by SpineAnalyzer® (QM SA) method, evaluates vertebral height loss that affects vertebral shape whereas the algorithm based qualitative (ABQ) method judges endplate integrity and includes guidelines for the differentiation of vertebral fracture and non- fracture deformities.

Results: Cross-sectional radiographs were assessed for 7,582 participants aged 45-95 years. With QM SA, the prevalence was 14.2% (95% CI: 13.4% to 15.0%), compared to 4.0% (95% CI: 3.6% to 4.5%) with ABQ. Inter-method agreement according to kappa (κ) was 0.24. The highest agreement between methods was among females (κ=0.31), participants aged above 80 (κ =0.40) and at the L1 level (κ =0.40). With ABQ, most fractures were found at the thoraco-lumbar junction (T12-L1) followed by the T7-T8 level, whereas with QM SA, most deformities were in the mid (T7-T8) and lower thoracic spine (T11-T12) with similar number of fractures in both peaks. Excluding mild deformities (grade 1 with QM) from the analysis increased the agreement between the methods from κ=0.24 to κ=0.40, whereas re-examining mild deformities based on endplate depression increased agreement from κ =0.24 to 0.50 (p-value< 0.001).

Conclusion: Vertebral fracture prevalence differs significantly between QM SA and ABQ; re-examining QM mild deformities based on endplate depression would increase the agreement

between methods. More wide-spread and consistent application of an optimal method may improve clinical care.

Introduction

Of all osteoporotic fractures, vertebral fractures are the most common type.[1] Vertebral fractures have been synonymous with the diagnosis of osteoporosis since its earliest description as a metabolic bone disorder.[2] Furthermore, osteoporotic vertebral fractures are a major health problem worldwide. Given the ageing of populations, osteoporotic vertebral fractures are likely to become an even increasingly important health issue. The costs of osteoporotic vertebral fractures were estimated to be € 1.5 billion in Europe in 2010[3] and are expected to have increased by more than 50% by 2025.[4]

Vertebral fractures may occur in the absence of trauma or after normal activities involving bending, lifting or turning.[1] Although, two thirds of vertebral fractures are not clinically detected, they are associated with decreased quality of life, back pain, functional limitations[5] and mortality[6] and can only be detected by formal screening. Vertebral fractures are often a first presentation of osteoporosis, therefore, accurate diagnosis is important to identify patients at high risk for future fractures. It has been shown that women with preexisting vertebral fractures have four times greater risk of subsequent vertebral fractures and 1.5 to 2 times greater risk of non-vertebral fractures than those without prior fractures, and this risk increases with the number and severity of prior vertebral fractures.[7-9] It is important to detect these fractures, since anti-osteoporotic therapy has been proven highly effective in reducing the risk of both non-vertebral and vertebral fractures.

Several methods for radiological assessment of vertebral fractures exist, but a gold standard is lacking.[10] The most commonly applied assessment methods include (semi-) quantitative morphometry (QM) and the algorithm based qualitative (ABQ) method. In contrast to semi-quantitative methods relying on expert visual inspection of height reduction, actual QM-based methods determine relative vertebral height loss by calculating ratios of the measured vertebral heights. Rather than only placing morphometry points manually on a

vertebral body, software packages such as Spine Analyzer® [11] apply Genant's classification[12] to define vertebral deformities. Finally, the algorithm based qualitative (ABQ) method by Jiang et al.[13] mainly judges endplate integrity, regardless of vertebral height reduction, and includes defined guidelines for the differentiation of vertebral fracture and non-fracture deformities. The key assumption is that the endplate is always deformed in vertebral fractures, and therefore endplate depression has perfect specificity for vertebral fracture. Vertebral height may appear to be decreased as a result of oblique image projection, specific diseases, and anatomical variants that can mimic vertebral fractures.[12-15] To deal with this misclassification, ABQ uses an algorithm to systematically rule out non-fracture deformities.

The aim of our study was to analyze differences in prevalence and fracture location between methods. We applied two methods, i.e., ABQ and SpineAnalyzer® software-assisted QM, for assessing vertebral fractures in the population-based Rotterdam Study, an ongoing prospective cohort study in elderly persons.

Materials and Methods

The Rotterdam Study: The Rotterdam Study is a prospective population-based cohort studying the determinants of chronic diseases and disability in Dutch men and women. Both the objectives and the study design have been described previously.[16] The study targets investigations on endocrine diseases like osteoporosis amongst others. It includes 14,926 inhabitants aged 45 years and over of Rotterdam city's Ommoord district in The Netherlands.

Vertebral fracture assessment: Radiographic examinations of the spine were obtained by a digitized Fuji FCR system (FUJIFILM Medical Systems). All radiographs were acquired according to a standardized protocol with a focus film distance of 120 cm. In some instances evaluability was suboptimal, mostly in the upper spine levels (supplementary Fig 1). In the

current report we have included participants with sufficient evaluability from T4-L4. Two teams, each composed of seven trained research assistants assessed lateral spine radiographs (T4-L4) independent of each other, using either ABQ or software-assisted QM (SpineAnalyzer®, Optasia Medical Ltd, Cheadle, UK). The mean inter- observer agreement for ABQ according to kappa statistic (κ) was moderate for both QM SA and ABQ (κ= 0.51 and κ=0.53 respectively). A subset of 76 radiographs were scored by two independent external readers; one reader with ABQ and one reader with QM SA; the agreement was poor, κ= 0.19. With ABQ, radiographs were triaged as normal, uncertain or definite fracture, based on integrity of the endplates. Definite and uncertain vertebral fractures were re-assessed by a musculoskeletal radiologist. SpineAnalyzer® software automatically identifies vertebral shape to calculate the exact heights of the vertebrae. After labeling the vertebrae of interest by placing thirteen points at the center of each vertebral body from L4 to T4, SpineAnalyzer® will place six morphometry points for each labeled vertebra, corresponding to the four corners and the middle of the vertebral body. The analyst can make manual adjustments to these six morphometry points to fine-tune their exact locations. The morphometry points are used to assess reductions in anterior, middle and posterior heights of the vertebrae by determining if one height measure is "reduced" in relation to another height (e.g., anterior height/posterior height<1 for a wedge shaped deformity). The SpineAnalyzer® software output provides a classification for deformities of shape (wedge, biconcave, crush) and severity (mild, moderate, severe). The wedge ratio is calculated by dividing anterior height by posterior height (hA/hP). Biconcavity is calculated by dividing mid height by posterior height (hM/hP). The calculation of crush fractures makes use of adjacent vertebral heights. Height loss less than 20% is considered normal. Mild fracture (grade one) is defined as height loss between ≥ 20% and<25%, moderate fracture (grade two) between ≥ 25% and<40% and

severe fracture (grade three) ≥ 40% according to Genant's classification scheme for osteoporotic vertebral fractures.[12]

Incident fracture were new fractures identified and reported by the general practitioners (GPs) or assessed from hospital records that occurred after baseline assessment. All events were then reviewed and coded by a research physician. For the current study we examined incident non-vertebral, hip and clinical-vertebral fractures.

Statistical analysis: We compared fracture prevalence and distribution according to vertebral level for QM SA and ABQ. Since there is no consensus whether most of the grade 1 or mild deformities are true osteoporotic vertebral fractures or not[14], we performed secondary analyses by excluding those fractures from the analysis. Agreement between the diagnostic approaches (inter- method agreement) and between raters (inter-rater agreement) for the identification of prevalent vertebral fractures was analyzed using kappa . The kappa value takes into account the proportion of agreement attributable to chance alone and can range from 0 (no agreement) to 1 (complete agreement); values greater than 0.8 are considered strong and values lower than 0.6 moderate[17]. Given that kappa is influenced by the imbalances in the distribution of marginal totals in the 2x2 table [18,19], together with kappa we have reported Bias Index (BI) which estimates the difference in proportions of "Yes" for the two raters, Prevalence Index which estimates the difference between the probability of "Yes" and the probability of "No" , observed agreement (p_o); proportion of positive agreement (p_{pos}) which estimates the conditional probability, given that one of the raters/ method, randomly selected, makes a positive rating, the other rater/ method will also do so; proportion of negative agreement (p_{neg}) which estimates the conditional probability, given that one of the raters/ method, randomly selected, makes a negative rating, the other rater/ method will also do so. We also calculated PABAK which is an index developed to account

for the effect that low prevalence and the difference in observer assessment of the frequency occurrence, have on kappa. All these statistics are derived from a 2x2 table as follows[19].

QM SA/ Rater 2 \ ABQ/ Rater 1	+	-
+	a	b
-	c	d

$P_o=(a+d)/N$ where N denotes total sample size

$P_e=(((a+b)(a+c))/N)+(((c+d)(b+d))/N))/N$

$P_{pos}=2a/(2a+b+c)$

$P_{neg}=2d/(2d+b+c)$

$BI=(b-c)/N$

$PI=(a-d)/N$

$PABAK=2P_o-1$

We calculated the above mentioned statistics per a) subject level; where prevalent cases were defined as subjects having at least one vertebra fractured from T4 to L4 and controls as having none of the vertebrae from T4 to L4 fractured, and per b) vertebral level; we counted as cases any fracture from T4 to L4; furthermore we calculated agreements of the methods between cohorts, sexes, age categories and vertebral level. We used four age categories: ≥ 45 and<60; ≥60 and<70; ≥ 70 and<80; ≥80. We separated vertebral level into three categories: T4-T9, T10-T12 and L1-L4. Additionally we assessed differences in baseline characteristics between cases and non-cases defined by either method and also differences between concordant and discordant cases defined as follows: QMSA + ABQ-, QM SA- ABQ+, QM SA+ ABQ+ against the reference group QM SA- ABQ- . The future incident fracture prediction ability by prevalent vertebral fractures scored by either method was estimated

using a Cox regression model adjusted for Age, Sex, BMI, cohort effect and FN-BMD with a mean follow up of 12 years. All analyses were performed using SPSS 21.0 (IBM Corp. NY, USA).

Results

Per subject analyses

Radiographs were assessed for 7,582 participants of which 61.7% (n=4,672) were from RS I, 21.8% (n=1,655) from RS II and 16.5% (n=1,255) from RS III. 60% of our study participants were females and age ranged from 46 to 95 years (mean 65.3, Fig. 1). QM SA scored vertebral fracture prevalence was 14.2% (95% CI: 13.4%-15.0%), compared to 4.0% (95% CI: 3.6%-4.5%) scored by ABQ. Participants who had sustained a fracture were significantly older according to both QM (67.4 vs. 64.9, p-value <0.001) and ABQ (70.4 vs. 65.1, p-value <0.001) compared to non-fractured participants. 54.5% of QM SA cases were females vs. 45.5 % males (p-value <0.001) and 74.0 % of ABQ cases were females against 26% males (p-value<0.001). Both QM SA and ABQ fractured participants had lower FN-BMD; 0.864 g/cm^2 vs. 0.890 g/cm^2 and 0.827 g/cm^2 vs. 0.894 g/cm^2, p-value <0.001 respectively. Fractured cases defined by ABQ were significantly shorter and lighter compared to the healthy participants 163.5 vs. 167.5 and 72.6 kg vs. 75.4 kg (p-value <0.001). No differences were seen between QM SA cases and controls in height and weight (p-value> 0.05) (Table 1a). When comparing (QM SA+) (ABQ-) participants vs. (QM SA-) (ABQ+), the latter had lower FN-BMD (0.846 vs. 0,877, p-value<0.001), were lighter (74.1 vs. 76.9, p-value<0.001), shorter (164.8 vs. 168.6) and comprised a higher number of females (74.3% vs. 50.1%, p-value<0.001) (Table1b). According to QM SA, the prevalence of vertebral fractures was higher among males compared to females (16.0% vs. 13.0%), whereas according to ABQ it was higher among females compared to males (5.0% vs. 2.6%) (Table 2). According to both methods the prevalence increased with increasing age (Table 3).

According to QM SA, 10% of the participants had only one spinal fracture, 2.6% had two fractures, 1.0% had three and 0.5% more than three fractures, whereas according to ABQ the estimates were lower with 2.9% of participants having only one fracture, 0.7% having two fractures, 0.2% three and close to 0% more than three.

The estimated concordance between ABQ and QM SA was $\kappa = 0.24$. When assessing agreement across sexes, it was significantly higher among females compared to males; $\kappa = 0.31$ vs. $\kappa = 0.14$, p-value<0.001 (Table 2). The agreement across age categories increased with increasing age; the highest kappa was among those aged above 80 and was significantly higher compared to the youngest group $\kappa = 0.40$ vs. 0.12, p-value<0.001 (Table 2).

Participants with a QM SA prevalent fracture had an increased risk for future non-vertebral fractures compared to those with absent prevalent vertebral fracture (HR= 1.15, 95% CI 1.007; 1.32) and also an increased risk of future clinical vertebral fracture (HR= 2.70, 95% CI 2.18; 3.35) but not for incident hip fracture (HR= 1.49, 95% CI 0.92; 1.71). The same trend was observed for participants with prevalent ABQ fractures although with higher estimates; participants with prevalent ABQ fracture had an increased risk to sustain a future non-vertebral fracture (HR= 1.30, 95% CI 1.06; 1.60), hip (HR= 1.47, 95% CI 1.05; 2.05) also an increased risk of incident clinical fractures (HR= 5.27, 95% CI 4.00; 6.77) compared to those with absent prevalent vertebral fracture (Fig 3).

Per vertebral body analyses

Among 7,582 participants, there were 1,574 (20.7%) vertebrae fractured according to QM SA and 447 (5.8%) according to ABQ. Figure 2 shows the distribution of osteoporotic vertebral fractures at each level assessed according to ABQ and QM SA. Both methods show a bimodal distribution, but according to ABQ, most fractures were found at the thoraco-lumbar junction (T12-L1) region, whereas according to QM SA, most deformities were at the middle (T7-T8) and lower thoracic regions (T11-T12), showing a more prominent bimodal

pattern (Fig. 2). The frequencies for QM SA deformities' classification of severity was 49.2% mild, 30.8% moderate and 4.7% severe; 53.5% of the deformities were wedge shaped, 11.9% were biconcave and 19.3% were crush (supplementary Table 1 and supplementary Figure 2). The agreement statistics per vertebral level could not be calculated for T4 since according to ABQ there were no T4 vertebrae fractured in any of the participants. The kappa statistic in the other vertebrae varied from 0.04 at T5 to 0.40 at L1. When assessing the agreement per region of the spine the highest agreement was in the L1-L4 region κ=0.37 (p-value<0.001) and when further stratifying by sex it reached κ=0.41 (p-value<0.001) among females (Table 4).

Excluding mild fractures from the study

We observed an increase in the net agreement between methods, mostly because the deformities with height loss but intact endplates were excluded. Out of 1,075 participants that were classified as fractured by QM SA, 614 of them had mild fractures. When excluding these subjects from the analysis, according to QM SA the prevalence decreased from 14.1% to 6.6 %. Excluding these participants slightly affected the prevalence of ABQ scored fractures with a decrease from 4.0% to 3.8%. On the other hand the kappa statistic increased from 0.24 to 0.40 (p-value<0.001) and reached its maximum among participants aged above 80, κ=0.47 among females κ=0.48 and at the L1 level κ=0.53 (Table 5). The prevalence of fractured vertebrae by grading of QM SA deformities is displayed by vertebral level distribution in Figure 4. According to QM SA, the highest concentration of fractured vertebrae was at T7-T8 and T11-T12-L1, showing again a bimodal distribution with almost the same number of fractured vertebrae for both peaks. A bimodal distribution was observed for ABQ as well, but with the highest peak at T12-L1.

Discussion

In this large population based study where we compared two assessment methods, osteoporotic vertebral fracture prevalence was four times higher when applying SpineAnalyzer® software-assisted QM compared to ABQ. Each method classified a considerable number of deformities that were assessed as normal by the other, reflected by poor between-method agreement statistics. Our study is the first to compare SpineAnalyzer® software-assisted QM and ABQ. According to ABQ, vertebral fracture prevalence was higher among females than males, whereas according to QM SA prevalence was higher among males. Differences in baseline characteristics were also observed; the difference in age, height, weight, FN-BMD and over-representation of females among cases compared to controls were stronger when they were defined by ABQ then when they were defined by QM SA. Also differences in BMD levels were observed among participants with discordant assessment of vertebral fractures, where participants with (ABQ+) (QM SA-) deformities had lower FN-BMD, weight and height compared to participants with (QM SA+) (ABQ-) deformities. We also observed difference in the ability to predict future non-vertebral and clinical vertebral fracture by prevalent vertebral fractures scored by either method with ABQ being more strongly associated with future fractures. The vertebral fracture prevalence estimate in our population for the ABQ method is similar to previous findings in other populations [13,20] mostly consisting of elderly females in a clinical setting; and also taking into account that we included subjects of both genders and even a subset comprising a relatively young population (RS-III). In previous work of the Rotterdam Study [21], including a sample of RS-I subjects assessed with the McCloskey-Kanis method [22], the prevalence was found to be 6.3%. This prevalence is intermediate between the prevalence of ABQ (~4.0%) and QMSA (~14.1%) and very similar to the prevalence of QM SA after excluding Grade 1 (~6.6%). The agreement was significantly higher in females compared to males, L1-L4 level and older age. The bimodal fracture distribution over the vertebral column was obvious for

the QM SA method in our cohort, with maxima at the mid-thoracic and lower thoracic regions including the thoraco-lumbar junction and less pronounced in ABQ. This pattern has been reported previously using other assessment methods. However, some argue that the more pronounced mid-thoracic peak with QM is to a great extent due to degenerative changes, normal anatomical variation (i.e., short vertebral height) and old traumatic fractures [23]. It has been put forward that ABQ would be able to differentiate these entities[15] compatible with our findings (Fig 3). When assessing QM SA morphometry, the far majority of deformities were classified as mild wedges located mostly at the T7-T8 level. By excluding QM SA-mild deformities, the difference in prevalence between the methods decreased and all agreement statistics increased.

We have assessed vertebral levels T4 to L4, as T1-T3 has poor evaluability and L5 is usually not affected by osteoporotic fractures. Several studies have compared assessment methods, but only a few have evaluated SpineAnalyzer® software or ABQ, and none have directly compared these two methods. SpineAnalyzer® software-assisted QM reading by a non-radiologist has been found to agree relatively well with conventional semi-quantitative (SQ) grading, i.e., visual estimation of vertebral body heights performed by experienced radiologists, with a kappa for agreement of 0.78.[24] ABQ comparisons with QM (Eastell-Melton and McCloskey definitions) have yielded kappa statistics between 0.39 and 0.64.[13] Most notably, the lowest agreement found to date is between ABQ and Genant's SQ methods, observing kappa statistics of 0.30 to 0.58.[15,25,26] The agreement between SpineAnalyzer software-assisted QM and ABQ in this study was even lower than the agreement between ABQ and Genant's SQ methods. This could have been further amplified because we have examined a relatively young and generally healthy population in RSIII, in which there might be many mild non-fracture deformities. This is also sustained by the results where kappa tended to increase with the increase of age. The kappa statistic is

associated with two paradoxes described by Feinstein et al.[18,19] These paradoxes arise from the chance-adjustment applied to kappa; adjustment that also helps to "standardize" and allow comparison across different studies. Kappa is estimated as the difference between *observed* and *expected agreement* divided by [*1 – expected agreement*]. Indeed in our study we observe a tendency towards *Paradox 1*, where there is high *expected agreement* (p_e) as well as high *observed agreement* which still results in a low kappa (Table 2). In addition, *Paradox 2* is also present given the population-based setting of our study, resulting in a large number of individuals without events, which creates unbalance of the marginal totals reflected in a high PI. The marginal totals are already determined by the (relatively low) prevalence of VFs and (healthy) population we studied and they can explain only partly the low kappa values. The remaining explanation of low kappa will arise from the method's separate performances for P_{pos} and P_{neg}. While kappa helps to compare agreement across studies, positive and negative agreement statistics help to better understand the individual study. In the present study, QM SA and ABQ agreed excellently to identify controls, but poorly to identify cases. Having said this and given that vertebral fracture diagnosis requires adaptation of current approaches to conciliate the differences between methods, we propose that one way would be by re-examining QM mild deformities for endplate depression. We simulated in our data a redistribution of the 2x2 table when reconsidering mild QM fractures for endplate depression and we saw that all agreement statistics increase significantly (supplementary Table 2c).

Nonetheless, it should be noted that agreement statistics concern precision of a study and may not necessarily relate to its validity. QM SA would not diagnose vertebral fractures in the case of endplate depression without reduced vertebral height, and conversely, ABQ would not diagnose a QM SA -based vertebral deformity with reduced height but intact endplates. More research is needed to clarify which of these discordant cases are clinically relevant vertebral fractures and which are false-positives.

It is important to recognize that although Spine Analyzer® software uses the Genant height criteria to judge severity of deformities defined by QM, QM methods on Spine Analyzer® software are *not* the same as the Genant semi-quantitative method[12]. While the Genant SQ method[12] unlike ABQ, does not specifically state how to differentiate non-fracture deformities from true fractures, it relies on the expertise of the evaluator [27] to discriminate them from vertebral height loss due to other causes such as degenerative remodeling and Scheuermann's disease [28]. In an accompanying article in this issue, Lentle et al.[29] employed the standard Genant methodology and draw similar conclusions with regard to the drastic differences in fracture prevalence and low concordance with a modified ABQ methodology.

Our overall aim was to objectively compare radiological assessment methods for osteoporotic vertebral fractures. Strengths of our study are that we systematically applied two very different assessment methods by two independent teams of trained readers which eliminates the risk of ascertainment bias. Applying two methods in a very large setting with two independent teams, proved to be very labor-intensive, requiring extra consensus meetings, supervision by musculoskeletal radiologists and double readings. Although radiographs were assessed by well-trained reader teams, it was not feasible to have all radiographs assessed by musculoskeletal radiologists. We are aware that more subtle endplate depression fractures could have been missed. As the Rotterdam Study is deemed representative of the general Dutch middle-aged to elderly population, we believe that our results may be extrapolated to other settings as well.

The semi-automated SpineAnalyzer® software-assisted QM method proved to be an excellent recording tool for research purposes, providing a standardized data output.[30] Surprisingly, ABQ was in our experience even more time-efficient, but this method requires more intensive initial training. Quantitative assessment is based on morphometry alone, which may result in the inclusion of deformities that are not truly vertebral fractures. For this

reason it might be better to refer to “deformities” instead of “fractures” for cases defined by QM. Yet, we experienced that further triage for both methods requires a lot of extra effort involving extra double-reading of up to thousands of participants. Further standardization and automation of this triage procedure with clear-cut classification criteria would be very helpful.

Vertebral fractures are often a first presentation of osteoporosis and should be regarded as an opportunity to trace individuals at high-risk for additional fractures and other related adverse health outcomes. To accomplish this, accurate vertebral fracture diagnosis is needed to identify these patients at high risk, as many effective treatment options are available. Conversely, individuals without true vertebral fractures should not be unnecessarily treated with medication , which is associated with unnecessary costs and potential adverse effects.[31] Improvement of radiological vertebral fracture definition, clearer criteria for non-fracture deformities differential diagnosis[32] and more wide-spread and consistent application of an optimal method may improve clinical care.

We have undertaken meticulous phenotyping on our ABQ and SpineAnalyzer® morphometric raw data. With these data, different cut-offs and vertebral fracture definitions could be linked to various clinically relevant outcomes. Furthermore, the remaining Rotterdam Study cohorts, which in total will yield ~11,000 subjects aged 45 years and over, will be assessed for the presence of osteoporotic vertebral fractures. In addition, our measurements could serve as population reference data.

In conclusion, we procured an impartial comparison of osteoporotic vertebral fracture assessment methods in the large population-based Rotterdam Study, with extensive recording of vertebral fracture distribution according to sex, age, deformity shape, severity and location. Osteoporotic vertebral fracture prevalence is significantly different when applying either software-assisted QM or ABQ. Further work is needed to reveal which of the discordant

cases are actually clinically relevant true vertebral fractures and which are not. We propose that mild deformities should be assessed for endplate depression, decreasing this way the false-positive QM fractures and conciliating the two methods.

Acknowledgements

We would like to thank Dr. Guirong Jiang for the training in the algorithm-based qualitative assessment method (ABQ). We are thankful to the employees from Optasia Medical Ltd who familiarized us with the use of the SpineAnalyzer® software. The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (014-93-015; RIDE2), RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study (particularly Lydia Buist and Hannie van den Boogert for acquisition of the radiographs) and the participating general practitioners and pharmacists. We thank René Vermeren. Nano Suwarno and Mart Rentmeester fort their technical support. Last but not least, we acknowledge the tremendous efforts from our team of radiographic readers.

Authors' contributions

LO, FK, SJB, MAI, EHGO, and FR designed the study. LO, FK, SJB, JBJvM, JHW, FJAvR collected and processed the data. LO, FK, SJB, JTS, EMC, JHW, FJAvR assessed and (statistically) analyzed the data. LO, FK, SJB, JTS, EMC, JBJvM, MCZ, GPK, EHGO, FR interpreted the results. LO, FK, SJB created the figures and tables. LO, FK, SJB, EHGO, FR drafted the manuscript. All authors (LO, FK, SJB, JTS, EMC, JBJvM, MAI, JHW, FJAvR, MCZ, GPK, EHGO, FR) read and revised the manuscript, and approved the final submitted version. LO and FK, EHGO and FR contributed equally. EHGO and FR assume

399 responsibility for the completeness and accuracy of the data and analyses, and for adherence
400 to the study protocol.
401

References

1. Szulc P, Bouxsein ML. Overview of osteoporosis: Epidemiology and clinical management. Vertebral Fracture Initiative Resource Document. 2011;PART I:1-65.
2. Cooper C. Epidemiology and public health impact of osteoporosis. Baillieres Clin Rheumatol. 1993;7(3):459-77.
3. Ström O, Borgström F, Kanis JA, et al. Osteoporosis: burden, health care provision and opportunities in the EU. Arch Osteoporos. 2011;DOI 10.1007/s11657-011-0060-1.
4. Burge R, Dawson-Hughes B, Solomon DH, Wong JB, King A, Tosteson A. Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025. J Bone Miner Res. 2007;22(3):465-75.
5. Nevitt MC, Ettinger B, Black DM, et al. The association of radiographically detected vertebral fractures with back pain and function: a prospective study. Ann Intern Med. 1998;128(10):793-800.
6. Bliuc D, Nguyen ND, Milch VE, Nguyen TV, Eisman JA, Center JR. Mortality risk associated with low-trauma osteoporotic fracture and subsequent fracture in men and women. Jama. 2009;301(5):513-21.
7. Klotzbuecher CM, Ross PD, Landsman PB, Abbott TA, 3rd, Berger M. Patients with prior fractures have an increased risk of future fractures: a summary of the literature and statistical synthesis. J Bone Miner Res. 2000;15(4):721-39.
8. Burger H, van Daele PL, Algra D, et al. Vertebral deformities as predictors of non-vertebral fractures. Bmj. 1994;309(6960):991-2.
9. Black DM, Arden NK, Palermo L, Pearson J, Cummings SR. Prevalent vertebral deformities predict hip fractures and new vertebral deformities but not wrist fractures. Study of Osteoporotic Fractures Research Group. J Bone Miner Res. 1999;14(5):821-8.
10. Oei L, Rivadeneira F, Ly F, et al. Review of radiological scoring methods of osteoporotic vertebral fractures for clinical and research settings. Eur Radiol. 2013;23(2):476-86.
11. Brett A, Miller CG, Hayes CW, et al. Development of a clinical workflow tool to enhance the detection of vertebral fractures: accuracy and precision evaluation. Spine (Phila Pa 1976). 2009;34(22):2437-43.
12. Genant HK, Wu CY, van Kuijk C, Nevitt MC. Vertebral fracture assessment using a semiquantitative technique. J Bone Miner Res. 1993;8(9):1137-48.
13. Jiang G, Eastell R, Barrington NA, Ferrar L. Comparison of methods for the visual identification of prevalent vertebral fracture in osteoporosis. Osteoporos Int. 2004;15(11):887-96.
14. Ferrar L, Jiang G, Adams J, Eastell R. Identification of vertebral fractures: an update. Osteoporos Int. 2005;16(7):717-28.
15. Ferrar L, Jiang G, Cawthon PM, et al. Identification of vertebral fracture and non-osteoporotic short vertebral height in men: the MrOS study. J Bone Miner Res. 2007;22(9):1434-41.
16. Hofman A, Brusselle GG, Darwish Murad S, et al. The Rotterdam Study: 2016 objectives and design update. Eur J Epidemiol. 2015;30(8):661-708.
17. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20(1):37-46.
18. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543-9.

19. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43(6):551-8.
20. Ferrar L, Roux C, Felsenberg D, Gluer CC, Eastell R. Association between incident and baseline vertebral fractures in European women: vertebral fracture assessment in the Osteoporosis and Ultrasound Study (OPUS). Osteoporos Int. 2012;23(1):59-65.
21. Van der Klift M, De Laet CE, McCloskey EV, Hofman A, Pols HA. The incidence of vertebral fractures in men and women: the Rotterdam Study. J Bone Miner Res. 2002;17(6):1051-6.
22. McCloskey EV, Spector TD, Eyres KS, et al. The assessment of vertebral deformity: a method for use in population studies and clinical trials. Osteoporos Int. 1993;3(3):138-47.
23. Adams JE, Lenchik L, Roux C, Genant HK. Radiological assessment of vertebral fracture. Vertebral Fracture Initiative Resource Document. 2011;PART II:1-48.
24. Kim YM, Demissie S, Genant HK, et al. Identification of prevalent vertebral fractures using CT lateral scout views: a comparison of semi-automated quantitative vertebral morphometry and radiologist semi-quantitative grading. Osteoporos Int. 2012;23(3):1007-16.
25. Ferrar L, Jiang G, Clowes JA, Peel NF, Eastell R. Comparison of densitometric and radiographic vertebral fracture assessment using the algorithm-based qualitative (ABQ) method in postmenopausal women at low and high risk of fracture. J Bone Miner Res. 2008;23(1):103-11.
26. Ferrar L, Jiang G, Schousboe JT, DeBold CR, Eastell R. Algorithm-based qualitative and semiquantitative identification of prevalent vertebral fracture: agreement between different readers, imaging modalities, and diagnostic approaches. J Bone Miner Res. 2008;23(3):417-24.
27. Grados F, Fechtenbaum J, Flipon E, Kolta S, Roux C, Fardellone P. Radiographic methods for evaluating osteoporotic vertebral fractures. Joint Bone Spine. 2009;76(3):241-7.
28. Armbrecht G, Felsenberg D, Ganswindt M, et al. Vertebral Scheuermann's disease in Europe: prevalence, geographic variation and radiological correlates in men and women aged 50 and over. Osteoporos Int. 2015;26(10):2509-19.
29. Lentle BC, Berger C, Probyn L, et al. Comparative Analysis of the Radiology of Osteoporotic Vertebral Fractures in Women and Men: Cross-Sectional and Longitudinal Observations from the Canadian Multicentre Osteoporosis Study (CaMos). In Press. 2017.
30. Oei L, Ly F, El Saddy S, et al. Multi-functionality of computer-aided quantitative vertebral fracture morphometry analyses. Quant Imaging Med Surg. 2013;3(5):249-55.
31. Breda SJ, Oei L, Oei EH, Zillikens MC. [Osteoporotic vertebral fractures or Scheuermann's disease?]. Ned Tijdschr Geneeskd. 2013;157(45):A6479.
32. Makurthou AA, Oei L, Saddy SE, et al. Scheuermann's Disease: Evaluation of Radiological Criteria and Population Prevalence. Spine (Phila Pa 1976). 2013.

Tables

Table 1a. Baseline characteristics of study participants shown by vertebral fracture status as scored by each definition. Fractured participants according to both QM SA and ABQ were significantly older, had lower FN-BMD and an over-representation of females. According to ABQ they were also shorter and lighter. Among QM SA cases, 57% were classified as grade 1, 37 % as grade 2 and 6% grade 3. Among ABQ defined cases, 39 were also scored as grade 1 by QM SA, 111 as grade 2 and 49 as grade 3.

		QM SA		ABQ	
	Overall N=7,582	Controls n=6,506	Cases n=1,076	Controls n=7,278	Cases n=304
Age	65.3 (8.8)	64.9 (8.6)	**67.4 (9.7)**	65.1 (8.7)	**70.4 (9.9)**
Sex (female)	4,516 (59.6)	3,930 (60.4)	**586 (54.5)**	4,291(59.0)	**225 (74.0)**
Height	167.4 (9.1)	167.4 (9.0)	167.5 (9.3)	167.6 (9.0)	**163.5 (8.5)**
Weight	75.3 (12.9)	75.2 (12.8)	76.0 (13.8)	75.4 (12.9)	**72.6 (13.4)**
BMI	26.8 (3.9)	26.8 (3.9)	27.0 (4.1)	26.8 (3.9)	27.1 (4.3)
FN- BMD*	0.890 (0.14)	0.895 (0.14)	**0.864 (0.14)**	0.894 (0.14)	**0.827 (0.14)**
QM SA Grade 1 2 3			 **614 (57.0)** **399 (37.0)** **63 (6.0)**		 **39** **111** **49**

***adjusted for age, sex, height, weight**

Table 1b. Baseline characteristics among participants with discordant and concordant assessment of vertebral fractures. Participants classified as cases according to QM but not according to ABQ were used as reference group for comparisons. Participants classified as cases according to ABQ but not to QM, were lighter, shorter , had lower FN-BMD and a higher representation of females.

N=7,582	(QM SA-) (ABQ-) (ref)	(QM SA+) (ABQ-)	(QM SA-) (ABQ+)	(QM SA+) (ABQ+)	(QM SA G2 or G3+) (ABQ+)
	N=(6,401)	N=(877)	N=(105)	N= (199)	N= (160)
Age	64.9 (8.5)	**66.4 (9.4)**	**67.6 (10.1)**	**71.9 (9.5)**	**72.4 (9.4)**
Sex (female)	3852 (60.2)	**439 (50.1)**	**78 (74.3)**	**143 (73.9)**	**121 (75.6)**
Height	167.4 (9.0)	**168.6 (9.1)**	**164.8 (8.0)**	**162.8 (8.7)**	**161.9 (8.4)**
Weight	75.27 (12.8)	**76.9 (13.7)**	**74.13 (13.2)**	**71.8 (13.5)**	**71.1 (13.0)**
BMI	26.8 (3.9)	27.0 (4.1)	27.2 (4.4)	27.0 (4.2)	27.0 (4.2)
FN- BMD*	0.896 (0.14)	**0.877 (0.14)**	**0.846 (0.14)**	**0.820 (0.14)**	**0.763 (0.14)**
QM SA Grade 1 2 3		 575 288 14		 39 111 49	 111 49

Table 2. Participants with prevalent vertebral fractures and agreement statistics between QM SA and ABQ, stratified by cohort and sex. The prevalence of VFs is the highest in RS III according to both QM SA and ABQ. The agreement statistics are the highest in RS I. According to ABQ, the prevalence of VFs is higher among females but not according to QM SA

	Cohort			Sex		Pooled
	RS I (N=4,672)	RS II (N=1,655)	RS III (N=1,255)	Males (N=3,066)	Females (N=4,516)	(N=7,582)
QM SA (%)	578 (12.4)	249 (15.0)	249 (19.8)	490 (16.0)	586 (12.9)	1076 (14.1)
ABQ (%)	190 (4.1)	59 (3.6)	55 (4.4)	79 (2.6)	225 (5.0)	304 (4.0)
Kappa	0.28	0.20	0.16	0.14	0.31	0.24
Observed agreement	0.89	0.86	0.81	0.85	0.89	0.87
Expected Agreement	0.85	0.82	0.77	0.82	0.83	0.83
Bias Index	0.08	0.11	0.15	0.13	0.08	0.10
Prevalence Index	-0.83	-0.81	-0.75	-0.81	-0.82	-0.81
Positive agreement	0.33	0.25	0.22	0.18	0.36	0.29
Negative agreement	0.94	0.92	0.89	0.91	0.94	0.93
PABAK	0.78	0.72	0.62	0.70	0.78	0.74

Table 3. Participants with prevalent vertebral fractures and agreement statistics between QM SA and ABQ stratified by age categories. The prevalence increases as age increases according to both methods. The highest prevalence is , as expected, among participants above 80 years old and kappa statistic is the highest in the same category.

	Age category			
	45-60 (N=2,396)	60-70 (N=2,932)	70 -80 (N=1,745)	>80 (N=509)
QM SA (%)	269 (11.2)	375 (12.8)	315 (18.1)	117 (23.0)
ABQ (%)	53 (2.2)	85 (2.9)	113 (6.5)	53 (10.4)
Kappa	0.12	0.20	0.30	0.40
Observed agreement	0.89	0.88	0.84	0.83
Expected agreement	0.87	0.85	0.77	0.71
Bias Index	0.09	0.10	0.11	0.12
Prevalence Index	-0.86	-0.84	-0.75	-0.66
Positive agreement	0.15	0.23	0.37	0.48
Negative agreement	0.94	0.93	0.91	0.90
PABAK	0.78	0.76	0.68	0.66

Table 4. Agreement statistics regarding number of fractured vertebrae by regions in the spine and by sex; note is per vertebral level. The lower in the spine is the fracture located, the higher is the agreement between methods.

	Spine Level								
N=7,582 Males n=3,066 Females n=4,516	T4-T9			T10-T12			L1-L4		
	Males	Females	Pooled	Males	Females	Pooled	Males	Females	Pooled
QM (%)	335(10.9)	339(7.5)	674(8.9)	156(5.1)	187(4.1)	343(4.5)	87(2.8)	129(2.9)	216(2.8)
ABQ (%)	29 (0.9)	51 (1.1)	80 (1.1)	24(0.8)	92 (2.0)	116(1.5)	43(1.4)	125(2.8)	168(2.2)
Kappa	0.10	0.17	0.14	0.14	0.39	0.29	0.28	0.41	0.37
Observed agreement	0.90	0.93	0.92	0.95	0.97	0.96	0.97	0.97	0.97
Expected Agreement	0.88	0.91	0.90	0.94	0.94	0.94	0.96	0.94	0.95
Bias Index	0.09	0.06	0.07	0.04	0.02	0.03	0.01	0.00	0.006
Prevalence Index	-0.88	-0.91	-0.90	-0.94	-0.94	-0.94	0.96	-0.94	-0.95
Positive agreement	0.12	0.18	0.15	0.16	0.40	0.31	0.29	0.43	0.38
Negative agreement	0.94	0.96	0.96	0.97	0.98	0.98	0.98	0.98	0.98
PABAK	0.80	0.86	0.84	0.90	0.92	0.92	0.94	0.94	0.94

Table 5. Agreement statistics regarding fractured subjects after excluding from the study those who had a mild fracture. After excluding participant with mild fractures from the study, all agreement statistics increase and the difference in prevalence between QM and ABQ decreases.

	Age Category				Sex		Pooled
	45-60 (N=2,217)	60-70 (N=2,698)	70 -80 (N=1,590)	>80 (N=463)	Males (N=2,768)	Females (N=4,200)	(N=6,968)
QM SA (%)	90 (4.0)	141 (5.2)	160 (10.0)	71 (15.3)	192 (6.9)	270 (11.2)	462 (6.6)
ABQ (%)	46 (2.0)	71 (2.6)	101 (6.3)	47 (10.1)	66 (2.4)	199 (4.7)	265 (3.8)
Kappa	0.25	0.35	0.47	0.53	0.28	0.49	0.41
Observed agreement	0.95	0.95	0.92	0.90	0.93	0.95	0.94
Expected Agreement	0.94	0.92	0.85	0.78	0.91	0.89	0.90
Bias Index	0.02	0.03	0.04	0.05	0.04	0.02	0.03
Prevalence Index	-0.94	-0.92	-0.83	-0.74	-0.90	-0.89	-0.89
Positive agreement	0.26	0.38	0.51	0.60	0.30	0.52	0.44
Negative agreement	0.98	0.97	0.96	0.94	0.97	0.97	0.97
PABAK	0.90	0.90	0.84	0.80	0.86	0.90	0.88

Figures

Fig. 1. Age at baseline distribution within the Rotterdam Study population, stratified by sex and cohort. RS III is the youngest cohort and RS I the oldest. Mean age among both sexes is 65.1 years but the study population is made up by approximately 60% females and 40% males.

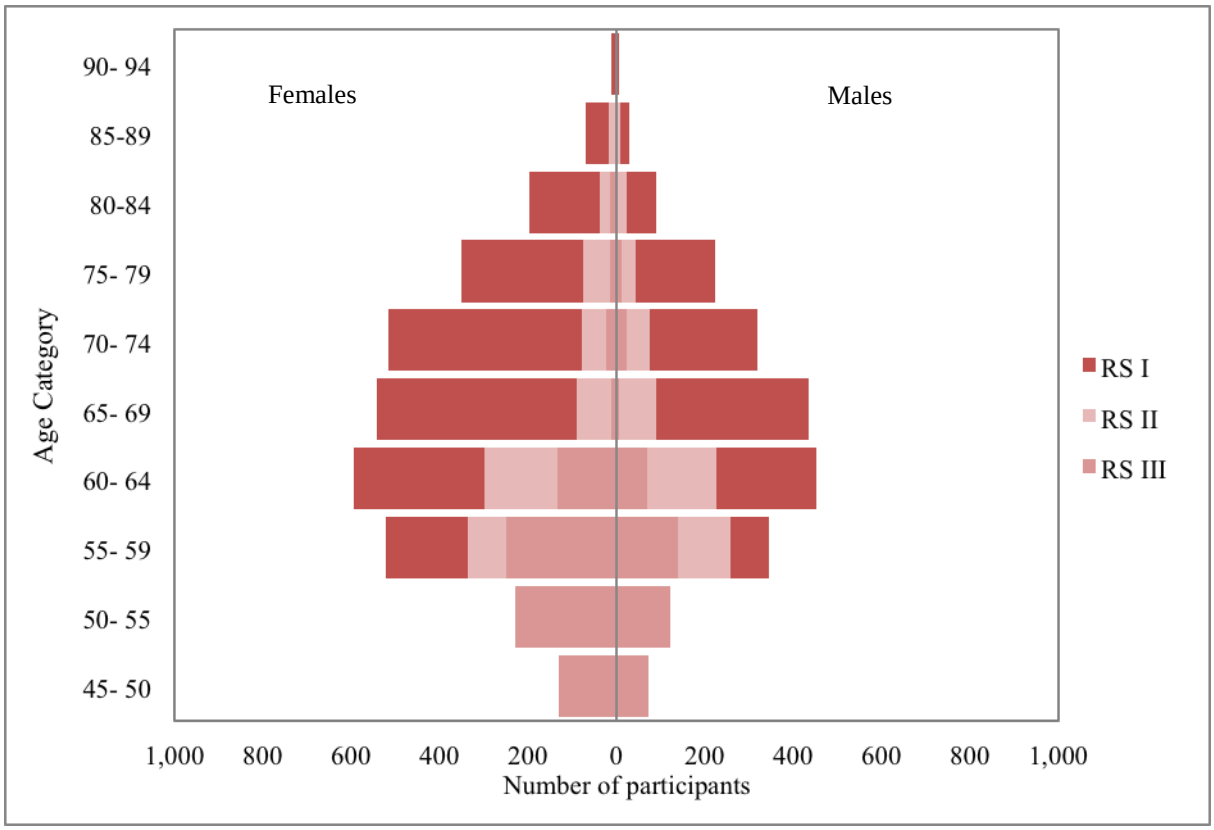


Fig. 2. Distribution of osteoporotic vertebral fractures across the thoracic and lumbar spine assessed according to the algorithm-based qualitative (ABQ) method and quantitative morphometry (QM) performed by SpineAnalyzer® software-assisted quantitative morphometry (vertebral height loss ≥ 20%). For both methods a bi-modal distribution can be seen but it is more pronounced for QM. According to QM the peaks are located at T7-T8 and T11-T12, whereas according to ABQ the highest peak is at T12-L1 and second highest at T7-T8.

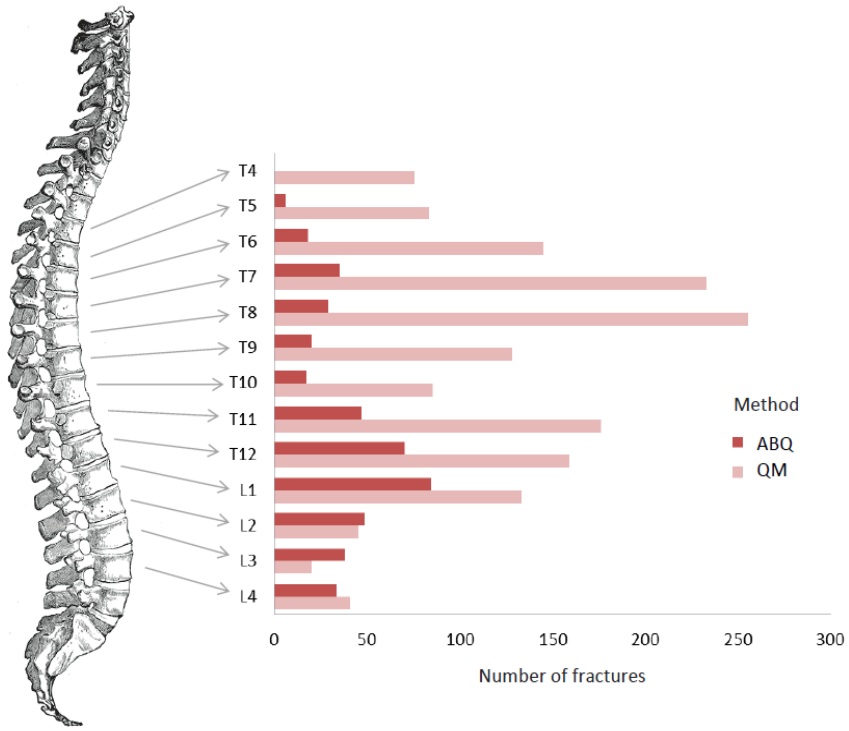
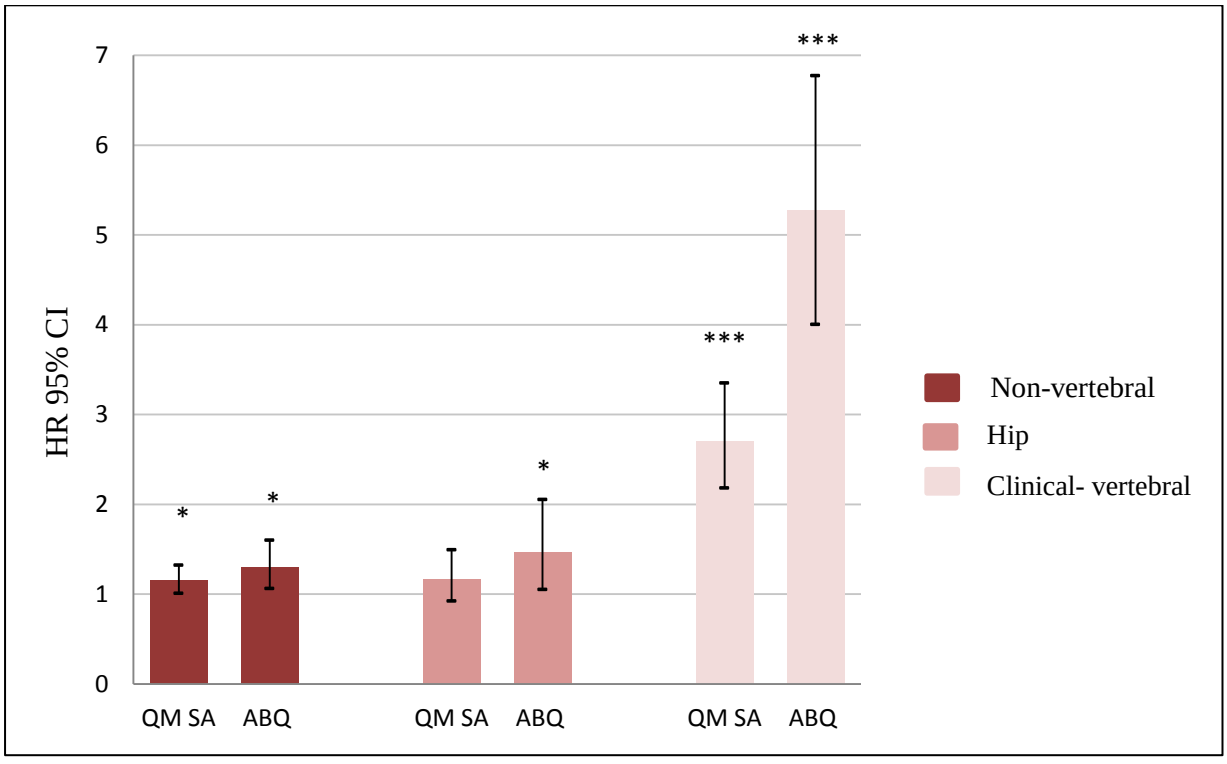
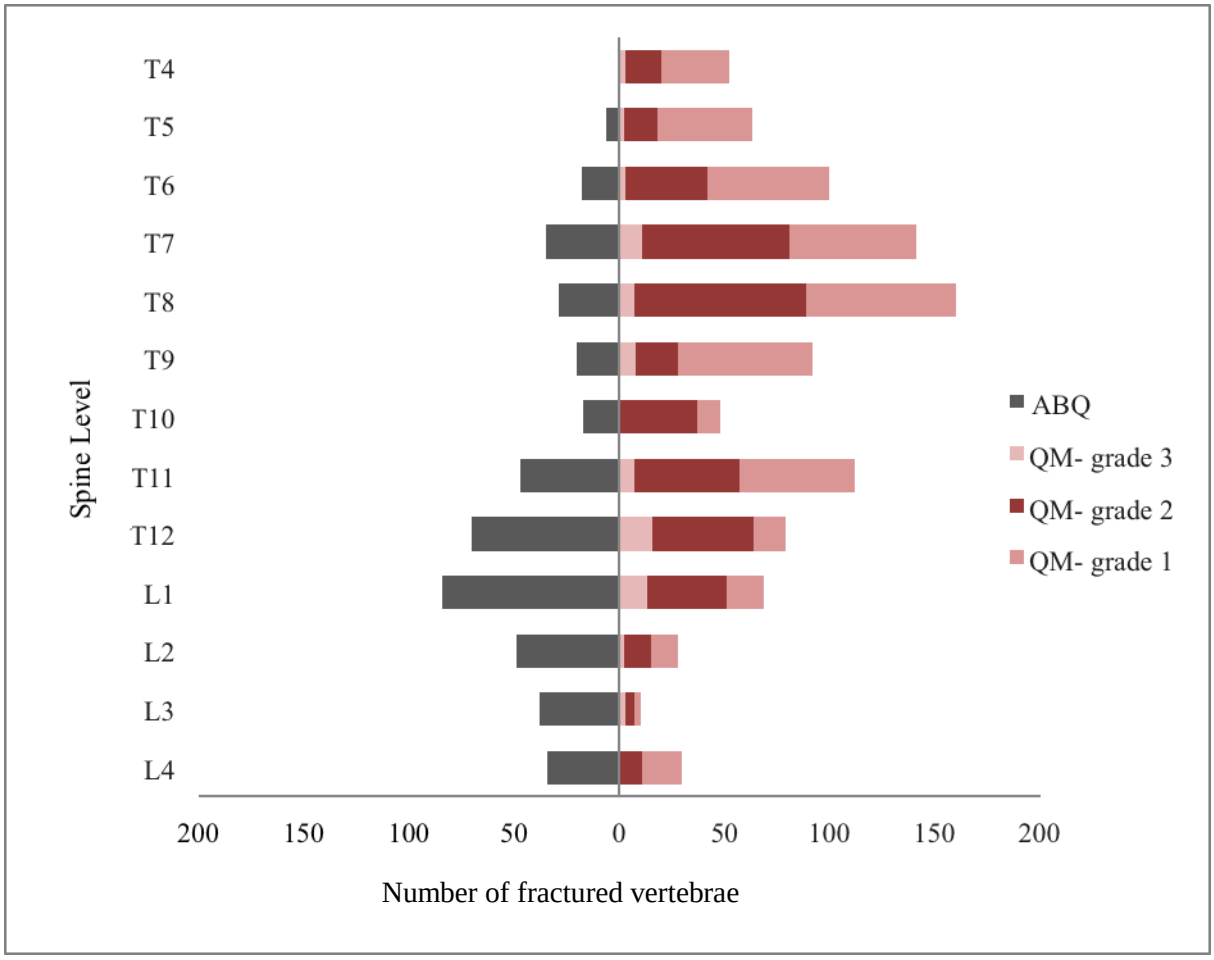


Fig. 3. The association between prevalent vertebral fractures scored by either method and incident non-vertebral and clinical vertebral fractures. During a mean follow-up time of 12 years, the 7,582 participants of this study sustained 1700 new non-vertebral fractures, 459 hip and 444 clinical-vertebral fractures. Participants with either prevalent QM or prevalent ABQ had increased risk of incident non-vertebral or clinical- vertebral fractures compared to participants who had not sustained either a QM or ABQ (respectively) fracture at baseline. Participants with an ABQ prevalent vertebral fracture at baseline were slightly more strongly associated with future non-vertebral fractures and significantly more strongly associated with incident clinical vertebral fractures compared to QM SA.



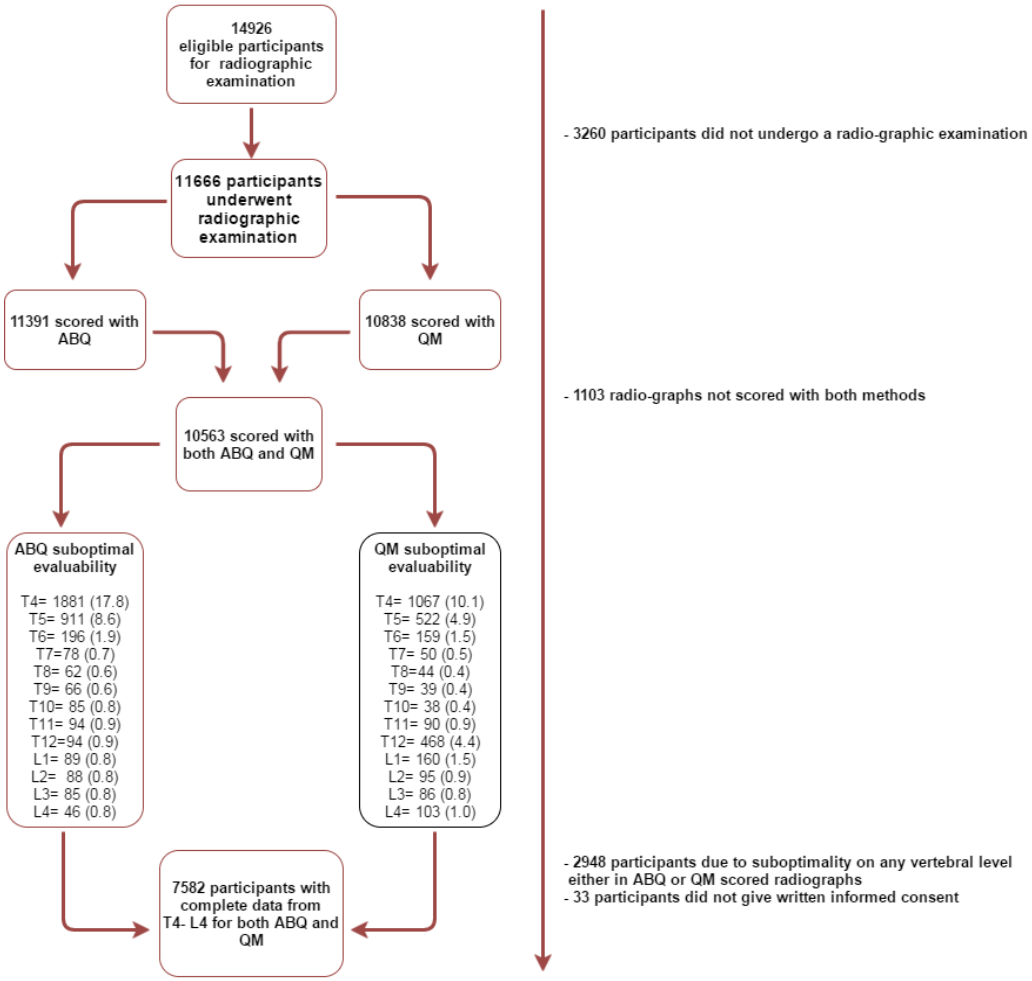
***p-value < 0.05**
*****p-value < 0.001**

Fig. 4. Distribution of osteoporotic vertebral fractures per vertebral level assessed with the algorithm-based qualitative (ABQ) method and quantitative morphometry (QM) performed by SpineAnalyzer® software-assisted quantitative morphometry. Mild deformities constitute around 62% of QM vertebral fractures, followed by grade two , 33% and the least common, grade three with 5%

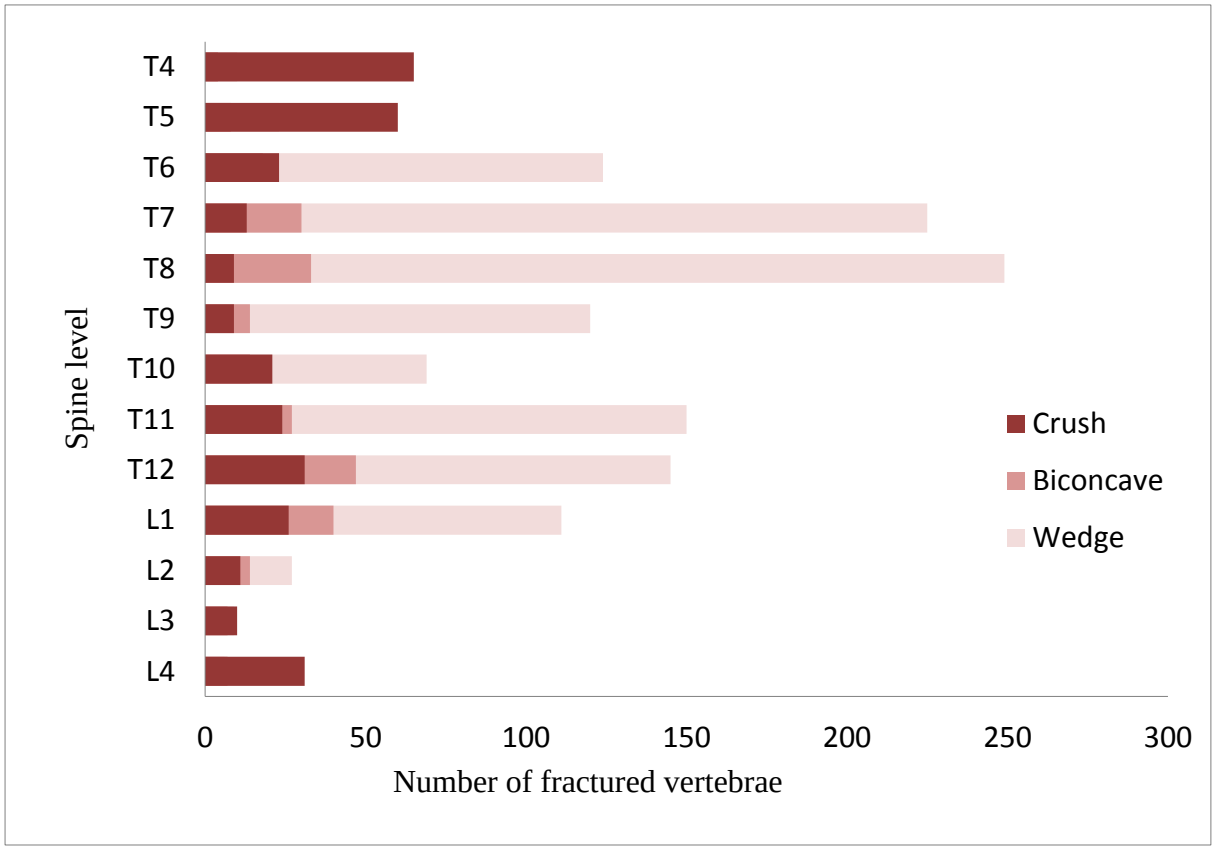


Supplementary Figures and tables

Supplementary Fig 1. Flowchart of the study participants. There were 14,926 participants that were eligible for radiographic examination and 3,260 did not undergo the exam. Out of 11,666 participants with radiography data, 828 were scored with ABQ but not with QM SA and 275 were scored by QM SA but not by ABQ, reducing the number of participants with radiographs scored by both methods to 10,563. Since we decided to perform analyses not only per subject level but also per vertebral body, we excluded participants that had any missing data from T4 to L4 level. Those missing were due to suboptimal visibility and no informed consent; this filter reduced the study population to 7,582 participants



Supplementary Fig 2. Distribution of QM fractures in the spine by morphometry. Crush fractures are mostly located at the upper thoracic level at T4-T5, biconcave at T7-T8 and T12-L1 and Wedge at T7-T8 and T11-T12.



Supplementary Table 1: Frequencies of QM SA vertebral fractures by shape and severity. In a population of 7,582 subjects, there were 1,574 vertebral bodies fractured of which 54.0% were wedge, 11.9% biconcave and 19.3% crush. On the other hand, 49.2% were classified as mild deformities, 30.8% as moderate deformities and 4.7% as severe

N=7,582 n=1,574	**Wedge** (n=842)	**Biconcave** (n=188)	**Crush** (n=304)
Mild- Grade 1 (n=775)	441	97	237
Moderate-Grade 2 (n=485)	348	73	64
Severe-Grade 3 (n=74)	53	18	3

Supplementary Table 2. The agreement between QM SA and ABQ and distribution in 2x2 tables in different scenarios; when applying the standard QM definition to QM SA, excluding mild deformities from the definition or assessing mild deformities based on endplate depression.

a) Agreement statistics for the study population when using the standard definition for QM SA

QM SA \ ABQ	+	-
+	199	877
-	105	6,401

	(N=7,582)
QM SA (%)	1,076 (14.2)
ABQ (%)	304 (4.0)
Kappa	0.24
Observed agreement	0.87
Expected Agreement	0.83
BI	0.10
PI	-0.81
Positive agreement	0.29
Negative agreement	0.93
PABAK	0.74

b) Agreement statistics for the study population when excluding subjects who had mild QM SA deformities

QM SA \ ABQ	+	-
+	160	302
-	105	6,401

	(N=6,968)
QM SA (%)	462 (6.6)
ABQ (%)	265 (3.8)
Kappa	0.41
Observed agreement	0.94
Expected agreement	0.90
Bias Index	0.03
Prevalence Index	-0.89
Positive agreement	0.44
Negative agreement	0.97
PABAK	0.88

c) Agreement statistics for the study population if we re-examine mild fractured subjects based on presence of endplate depression. Out of 614 subjects that had a mild fracture, 39 were classified as fractured also by ABQ. If we classify those 39 mild+ and ABQ + as true positives and the 575 remaining we classify as true negatives, the redistributed 2x2 table would look like the one below. Calculating agreement statistics for that 2x2 table, produces even higher agreement than just excluding those deformities from the study analysis.

		ABQ	
		+	-
QM SA	+	199	302
	-	105	6,976

	(N=7,582)
QM SA (%)	501 (6.6)
ABQ (%)	304 (4.0)
Kappa	0.50
Observed agreement	0.95
Expected Agreement	0.90
Bias Index	0.03
Prevalence Index	-0.89
Positive agreement	0.50
Negative agreement	0.97
PABAK	0.90